WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Credit Score Classification

Preethi Elagonda[1], Dr. M. Ramchander[2], Dr. G.N.R. Prasad[3]

MCA III Semester Student, CBIT(A), Hyderabad, India,
Email: preethielagonda555@gmail.com
Dr. M. Ramchander - Assistant Professor, CBIT(A), Hyderabad, India,
Email: mramchander_mca@cbit.ac.in
Dr. G.N.R. Prasad - Sr. Assistant Professor, CBIT(A), Hyderabad, India,
Email: gnrprasad_mca@cbit.ac.in

ABSTRACT :

Assessing the financial reliability of individuals is essential for lending organizations, as it plays a significant role in loan approvals and managing risk. This research presents a machine learning framework that classifies credit scores into three categories: Good, Standard, and Poor. Key financial factors, including income, repayment habits, outstanding debts, and loan histories, were examined following comprehensive preprocessing utilizing Python-based tools. Two supervised learning algorithms, Logistic Regression and Random Forest, were implemented and compared. The experimental results indicated that Random Forest attained higher accuracy, demonstrating that it is a more effective model for predicting credit risk. The created system reduces dependence on manual evaluations, improves the efficiency of financial decision-making, and exhibits considerable potential for incorporation into real-time banking systems.

Keywords — Credit Scoring, Financial Risk Prediction, Machine Learning Models, Random Forest, Logistic Regression, Automated Classification

Introduction

Evaluating creditworthiness is a key procedure in the financial sector, enabling banks and lending institutions to ascertain an individual's eligibility for loans or other financial products. Historically, this evaluation relied on manual record assessments, which were not only labor-intensive but also susceptible to inconsistencies and errors from human intervention. With the swift expansion of digital financial information, machine learning has emerged as an effective method for automating the evaluation of credit risk and enhancing reliability.

A credit score acts as a reflection of a person's past financial behavior, encompassing repayment habits, income levels, existing debts, and the number of active credit accounts. Effectively categorizing these scores into classifications of Good, Standard, or Poor helps organizations mitigate the risk of loan defaults and make more equitable lending choices. Machine learning algorithms, by detecting patterns within this data, yield more precise and consistent predictions than traditional assessment approaches.

This research introduces a machine learning framework aimed at the classification of credit scores. The study included data preprocessing, encoding of categorical variables, and exploratory analysis to derive significant insights. Two supervised learning methodologies—Logistic Regression and Random Forest—were utilized, and their performance levels were contrasted. The results show that Random Forest achieved superior accuracy, validating its capability to manage a wide array of financial attributes and categorical data.

The findings highlight that automated credit scoring can greatly improve both efficiency and reliability while providing scalability for practical financial systems. Additionally, the suggested approach has the potential for further development by integrating advanced ensemble models or adapting it for real-time implementation within banking contexts.

Objectives

The main goals of this project are:

1. Automated Credit Score Categorization – To create a system that classifies individuals into Good, Standard, or Poor categories by utilizing machine learning methods.
2. Data-Driven Decision Making – To substitute manual evaluations with an approach based on data that improves fairness, consistency, and efficiency in assessing credit risk.
3. Model Implementation and Comparison – To put into practice Logistic Regression and Random Forest classifiers, and assess their performance on financial datasets.
4. Performance Evaluation – To evaluate the efficiency of the models with standard metrics such as accuracy, precision, and recall.
5. Practical Application – To illustrate how the system developed can help financial institutions minimize loan default risks and enhance decision-making processes.

6. Scope for Enhancement – To lay the groundwork for future improvements, including the integration of advanced ensemble techniques, additional features, and real-time deployment.

METHODOLOGY

The approach taken for this project included the following phases:

1. **Requirement Analysis –** Recognized the necessity for an automated solution to categorize credit scores and minimize manual assessment in financial decision-making.
2. **Data Preprocessing –** Cleaned the dataset, managed missing values, and transformed categorical variables like Credit Mix and Occupation into a usable format.
3. **Model Development –** Employed Logistic Regression and Random Forest algorithms utilizing Python and Scikit-learn.
4. **Training & Testing –** Divided the dataset into training (80%) and testing (20%) subsets to assess the performance of the models.
5. **Performance Evaluation –** Evaluated the models based on accuracy metrics, with Random Forest yielding better results.
6. **User Prediction Module –** Created a console-based application that takes financial information as input and forecasts the credit score category along with probability outcomes.

System Design / Architecture:

The architecture of the credit score classification system was developed with a modular approach to maintain clarity, scalability, and ease of implementation. The primary components consist of:

1. **Input Layer –** Gathers financial information such as earnings, existing debt, the number of loans, instances of late payments, and credit mix.
2. **Preprocessing Module –** Manages missing data, applies label encoding for categorical variables, and normalizes the input data to ensure consistency.
3. **Feature Selection –** Determines critical attributes (such as earnings, late payments, existing debt) that impact credit score classification.
4. **Model Training Module** – Applies Logistic Regression and Random Forest classifiers to the training dataset.
5. **Prediction Engine** – Utilizes the trained Random Forest model to categorize applicants as Good, Standard, or Poor.
6. **Evaluation Module** – Assesses model accuracy and provides probability-based outputs for each credit score category.
7. **User Interaction Layer** – Features a console-based interface where users can input information and obtain predicted results.

Implementation & Results

Implementation

The system was developed using Python along with several tools:

- **Libraries:** Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn
- **Algorithms:** Logistic Regression and Random Forest
- **Environment:** Jupyter Notebook / Google Colab
- **Dataset:** credit_score.csv which includes financial metrics such as income, loans, payment delays, and outstanding debt

The steps undertaken were:

1. Loaded and analyzed the dataset.
2. Processed the data by addressing missing values and converting categorical variables into numerical format.
3. Divided the dataset into a training set (80%) and a testing set (20%).
4. Trained both Logistic Regression and Random Forest algorithms.
5. Assessed the models using accuracy metrics and probability outputs.
6. Created a straightforward console-based interface for user predictions.

Results

- Logistic Regression obtained moderate accuracy, but struggled with complex relationships.
- Random Forest demonstrated greater accuracy and stability, making it the model of choice.
- Example predictions:
- Poor Credit – Low income, high debt → Predicted “Poor”
- Standard Credit – Moderate income, few loans → Predicted “Standard”
- Good Credit – High income, low debt, no delays → Predicted “Good”
    - The final accuracy of the Random Forest model proved to be significantly better than that of the Logistic Regression model, confirming its reliability for practical applications.

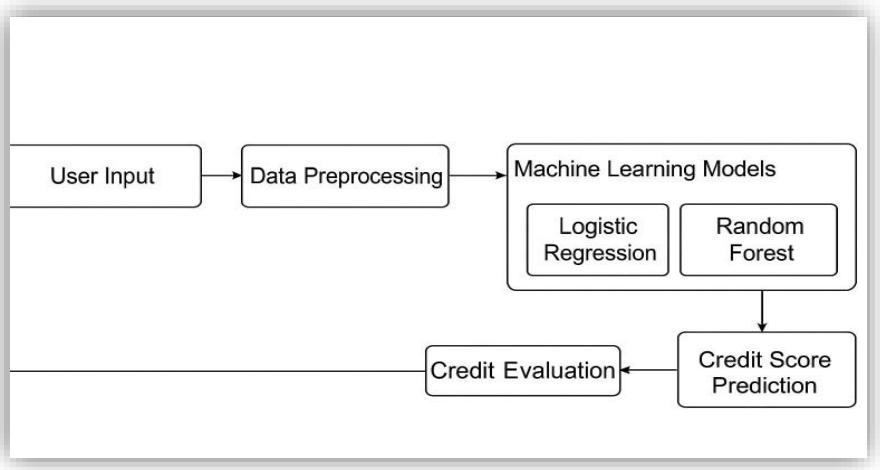

Figure 1: This image is a flow diagram showing the process of Credit Score Prediction using Machine Learning.

```
Random Forest Model Accuracy: 81.63 %

Credit Score Prediction:
Annual Income: 100000
Monthly Inhand Salary: 8000
Number of Bank Accounts: 5
Number of Credit cards: 5
Interest rate: 5
Number of Loans: 1
Average number of days delayed: 0
Number of delayed payments: 0
Credit Mix (Bad: 0, Standard: 1, Good: 2): 2
Outstanding Debt: 500
Credit History Age (in days): 500
Monthly Balance: 1500

Predicted Credit Score = Good

Prediction Probabilities:
  Probability of Good: 0.42
  Probability of Poor: 0.17
  Probability of Standard: 0.41
```

Figure 2: Prediction results from a credit scoring model. For the given data, the predicted score is "Good" with a confidence probability of 42%.

```
Random Forest Model Accuracy: 81.63 %

Credit Score Prediction :
Annual Income: 30000
Monthly Inhand Salary: 2500
Number of Bank Accounts: 1
Number of Credit cards: 1
Interest rate: 12
Number of Loans: 1
Average number of days delayed by the person: 2
Number of delayed payments: 1
Credit Mix (Bad: 0, Standard: 1, Good: 2): 1
Outstanding Debt: 5000
Credit History Age (in years): 5
Monthly Balance: 2500

Predicted Credit Score = Standard
Probability of Good: 0.37
Probability of Poor: 0.25
Probability of Standard: 0.38
```

Figure 3: Prediction results from a credit scoring model. For the given data, the predicted score is "Standard" with a confidence probability of 38%.

Conclusion

This research showcased the application of machine learning for the classification of credit scores in an automated manner. By examining financial factors such as income, borrowing history, existing debts, and payment delays, the system was able to categorize individuals into Good, Standard, and Poor classifications. Among the models applied, Random Forest achieved superior performance compared to Logistic Regression, delivering greater accuracy and dependability. This study underscores the potential of data-driven methodologies to minimize manual workload, enhance fairness, and improve decision-making within financial organizations.

Future Work

1. Integration of more sophisticated models like Gradient Boosting or Deep Learning to further boost predictive accuracy.
2. Addition of further features such as transaction history, employment information, and spending habits.
3. Creation of a web or mobile application for real-time interaction with users.
4. Implementation of the system within financial institutions to facilitate large-scale automated assessments of credit risk.
5. Investigation of explainable AI approaches to improve transparency in the decision-making process.

REFERENCES

[1] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed. O'Reilly Media, 2019.
[2] S. Raschka and V. Mirjalili, Python Machine Learning, 3rd ed. Packt Publishing, 2019.
[3] S. Lessmann, B. Baesens, H. Seow, and L. Thomas, "A comparative analysis of classification algorithms for credit scoring," European Journal of Operational Research, vol. 247, no. 1, pp. 124–136, 2015.
[4] Scikit-learn Documentation, "Machine Learning in Python," [Online]. Accessed: https://scikit-learn.org
[5] Pandas Documentation, "Python Data Analysis Library," [Online]. Accessed: https://pandas.pydata.org
[6] Kaggle Datasets, "Credit Score Classification Dataset," [Online]. Accessed: https://www.kaggle.com